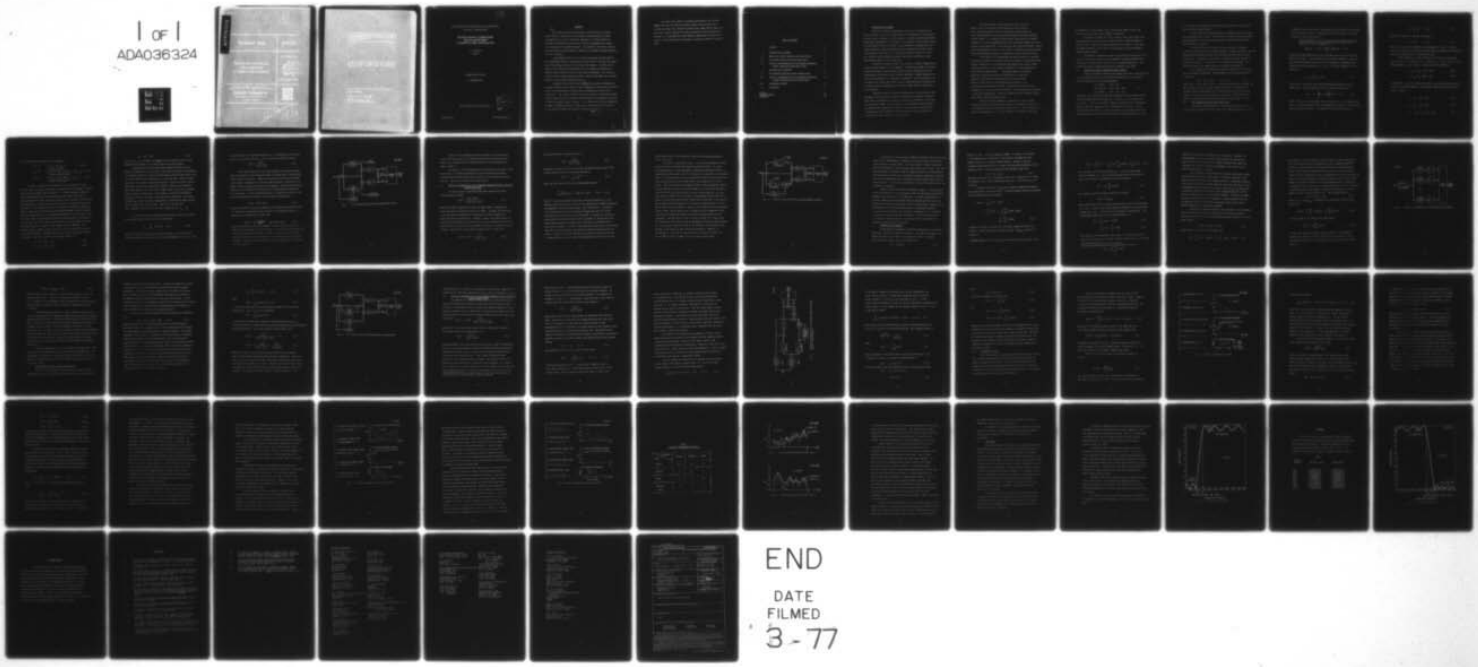
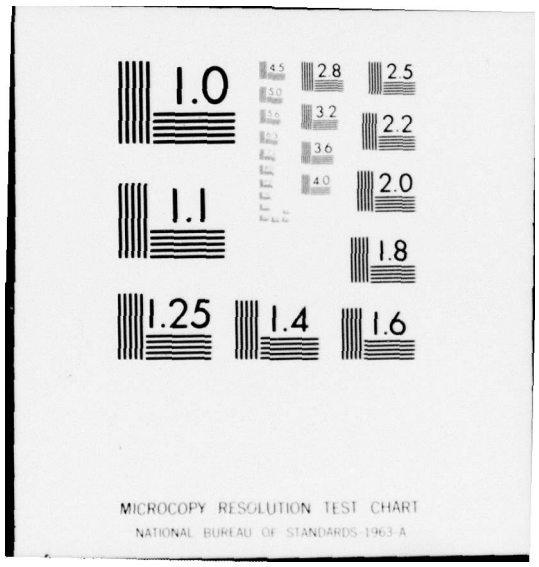
AD-A036 324 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 9/4
OPTIMUM SPEECH CLASSIFICATION AND ITS APPLICATION TO ADAPTIVE N--ETC(U)
NOV 76 R J MCAULAY F19628-76-C-0002
UNCLASSIFIED TN-1976-39 ESD-TR-76-320 NL

1 OF 1
ADA036324



END

DATE FILMED
3-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Technical Note

1976

R. J. M

Optimum Speech Classification and Its Application to Adaptive Noise Cancellation

9 Novem

Prepared for the Defense Communications Agency
under Electronic Systems Division Contract F19628-76-C-0002 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, for the Military Satellite Office of the Defense Communications Agency under Air Force Contract F19628-76-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

(12)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

OPTIMUM SPEECH CLASSIFICATION AND ITS APPLICATION TO ADAPTIVE NOISE CANCELLATION

R. J. McAULAY

Group 24

TECHNICAL NOTE 1976-39

9 NOVEMBER 1976

Approved for public release; distribution unlimited.

LEXINGTON MASSACHUSETTS

ABSTRACT

The problem of determining whether a given interval of a speech signal should be classified as voiced speech, unvoiced speech or silence is formulated as a test of statistical hypotheses. A robust detector is obtained by modelling the speech and the acoustic background noise signals as correlated Gaussian random processes. The methods of statistical decision theory are applied to these models to synthesize an optimum, minimum probability of error, classifier.

The optimum classifier is an estimator-correlator receiver which is well approximated using a linear phase high pass filter in the unvoiced channel and a linear phase low pass filter in the voiced channel. A clutter filter appears in the reference channel which tries to eliminate as much noise as possible before forming the unvoiced and voiced correlations. The statistics of the noise are learned during the silent intervals which makes the classifier adaptive to time-varying noise statistics.

Knowledge of the clutter correlation function permits implementation of adaptive Wiener filters which are used to eliminate as much noise as possible prior to the determination of pitch and the estimation of the LPC filter coefficients. The clutter filtered voiced speech signal is then passed through a bank of comb filters and the pitch estimate chosen to correspond to the filter for which the output energy is largest. It is shown that this pitch estimation strategy is optimum and robust as long as the correlation time of the noise is less than the minimum pitch period of interest.

The robust LPC vocoder is evaluated experimentally for Airborne Command Post noise for which the unvoiced speech signal-to-noise ratio is often less than 0 dB. Based on listening tests comparing the input speech plus noise, versus standard LPC synthesis techniques versus the robust LPC vocoder, it is concluded that rather dramatic improvements in speech intelligibility can be obtained at the expense of a marginal increase in computation time.

TABLE OF CONTENTS

I. INTRODUCTION AND SUMMARY

There are a variety of applications in which it is necessary to be able to classify a given set of speech data as corresponding to voiced speech, unvoiced speech or silence. For the synthesis of speech using Linear Predictive Coding (LPC) techniques[1-4], for example, it is necessary that the speech signal be classified as voiced or unvoiced. This information is transmitted to the speech synthesizer along with coefficients that represent an all-pole linear filter model for the vocal tract. For voiced speech the filter is excited by a periodic train of impulses, whereas a white noise excitation is used when unvoiced speech is to be synthesized.

The ability to detect silence is of interest in digital communications in which channel capacity is at a premium[5]. By detecting intervals of silence, other data streams can be interleaved with the speech conversation thereby maximizing the utilization of the available bandwidth. Another application of silence detection arises in conferencing situations[5]. By detecting when a set of speakers are silent, their lines can be disconnected from the superposition of inputs so that an enhancement of synthesizer input signal-to-noise ratio can be obtained.

Solutions to the classification problem have, for the most part, been developed on an ad hoc basis in which an individual discriminant is proposed which seems to characterize in one way or another the attributes of the three possible speech events. In a recent paper, Atal and Rabiner[6] have proposed an algorithm that simultaneously computes five of the most significant discriminants and uses a hypothesis testing strategy to assign a given set of observations to one of the three speech classes.

With few exceptions, most notably the work of Atal and Rabiner, most of the speech research reported to date has dealt with a speech environment that has been carefully controlled in the sense that background noise and interference signals have been eliminated from the speech. It is generally known that the intelligibility of modern vocoders is seriously degraded when noise and interference signals are superimposed on the speech data[5]. Since there are many practical problems in which noise and interference arise, it is of interest to develop more general speech processing techniques designed to eliminate the noise as much as possible.

In this paper it is assumed that the speech signals are corrupted by additive Gaussian noise that may or may not be white. The unvoiced speech signal is modelled as a zero mean Gaussian random process having a known covariance function. Voiced speech is modelled as a zero mean Gaussian quasi-periodic random process. Using these models as a starting point the classification problem is formulated as a statistical hypothesis test and solved using statistical decision theory. Subject to the validity of the underlying speech models, the resulting signal processing algorithm is optimum in the sense that the probability of a decision error is minimized. The advantage of this approach is that the discrimination criteria are synthesized from the model, rather than being selected on an ad hoc basis.

The classification problem is recognized as a Gauss-in-Gauss detection problem for which solutions have been catalogued by Van Trees[7]. The estimator-correlator structure was chosen since it led most naturally to a practical implementation. If pitch information is available, additional

discrimination can be provided in the voiced speech channel using a comb filter tuned to the most recent estimate of the pitch.

The ability to detect the silent intervals (noise alone) means that the statistics of the clutter can be learned and used to implement adaptive Wiener filters to enhance the speech signals prior to coding. In this mode the adaptive prefilter can be used as a pre-processor for any narrowband or wideband speech encoder.

An extensive experimental program was developed to evaluate the classifier in a variety of acoustic noise environments including shipboard noise, office noise, helicopter noise and noise in an airborne command post. The results for airborne command post noise are included in this paper.

II. MODELS FOR SILENCE, UNVOICED AND VOICED SPEECH

The basic problem of detecting the presence of silence, unvoiced speech or voiced speech in a given set of data can be formulated as a statistical test for choosing one of the three hypotheses:

$$\begin{aligned} H_1&: \text{silence:} \quad y(n) = w(n) \\ H_2&: \text{unvoiced:} \quad y(n) = u(n) + w(n) \\ H_3&: \text{voiced:} \quad y(n) = v(n) + w(n) \end{aligned} \tag{2.1}$$

where $w(n)$, $u(n)$ and $v(n)$ represent the nth sample of noise, unvoiced speech and voiced speech waveforms respectively. Based on a set of observations $y(1), y(2),\ldots,y(N)$ it is desired to develop a decision rule for determining which of the three hypotheses "best" characterizes the data set. This is the classification problem. In order to synthesize an optimum decision rule in the sense that a classification is made with minimum probability of error,

it is necessary to develop statistical models that characterize the data for each of the three speech events.

To begin with, the interference will be assumed to consist of simply zero mean white Gaussian noise. Once the detector structure has been analyzed and understood for this case the generalization to non-white noise spectra follows almost by inspection.

In order to derive the structure of the classifier it suffices to model the unvoiced and voiced speech waveforms as sample functions of Gaussian random processes having zero means and covariance functions $R_u(k)$ and $R_v(k)$ respectively. In addition voiced speech is assumed to be quasi-periodic in the sense that $R_v(k+T) = R_v(k)$ where T is the period of the process. This means that almost every sample function is periodic with period T[8].

The preceding discussion can be summarized succinctly by the following set of modelling equations. Under hypothesis H_i the observed data set is given by:

$$y(n) = s_i(n) + w(n) \qquad i = 1, 2, 3 \qquad (2.2)$$

where $s_1(n) = 0$ for silence, $s_2(n)$ is a Gaussian random process with mean zero and covariance $R_u(k)$ for unvoiced speech and $s_3(n)$ is a zero mean quasi-periodic Gaussian random process with covariance function $R_v(k)$ for voiced speech. In all cases the noise term $w(n)$ represents a zero mean Gaussian white noise random process having the correlation function $R_w(k) = \sigma^2 \delta(k)$.

III. THE OPTIMUM CLASSIFIER AGAINST WHITE NOISE

The optimum classifier processes the raw speech data $y(1), y(2), \ldots y(N)$ in such a way that a decision is made with minimum probability of error

on whether the given interval of signal should be classified as voiced speech, unvoiced speech or silence. Using statistical decision theory the minimum probability error decision rule is:

> "Declare hypothesis H_i to be true if and only if the a posteriori probability that H_i is true conditioned on the observation set $y(1), y(2),...y(N)$ is largest," i.e.,

$$p[H_i|y(N),...,y(1)] = \max_{k=1,2,3} p[H_k|y(N),...,y(1)]$$

Signal processing configurations of the likelihood ratio test have been documented by Van Trees[7]. For the special case of ternary hypotheses, zero means and stationary random processes the test is implemented by computing three sufficient statistics denoted by ℓ_i, i=1,2,3. The first component of the ith statistic is

$$\ell_{yi} = \sum_{n=1}^{N} y(n)\ \hat{s}_i(n) \tag{3.1}$$

where $\hat{s}_i(n)$ is the linear least squares unrealizable estimate of the ith signal $s_i(n)$. The bias component of the ith sufficient statistic is

$$\ell_{Bi} = -\frac{T}{2} \int_{-\infty}^{\infty} \ell n\ [1 + \frac{G_i(f)}{N_o/2}]\ df \quad i=1,2,3 \tag{3.2}$$

where $T = N/F_s$ is the observation time of the process, F_s is the sampling rate, $G_i(f)$ is the power spectrum of the ith random process and $N_o/2$ is the two-sided white noise spectral density. The complete ith sufficient statistic is

$$\ell_i = \ell_{yi} + \ell_{Bi} \qquad i=1,2,3 \tag{3.3}$$

and the test consists of choosing the largest of

$$\ell_i + \ell n\, P_i \qquad i=1,2,3 \tag{3.4}$$

where P_i is the a priori probability that hypothesis H_i is true. The goal now is to use the gross attributes of speech signals to simplify the computations involved in implementing the likelihood ratio test.

Under hypothesis H_1, which corresponds to silence, the anticipated signal is $s_1(n) = 0$. Therefore $\hat{s}_1(n) = 0$ whence $\ell_{y_1} = 0$, $\ell_{B_1} = 0$ and $\ell_1 = \ell nP_1$. The likelihood ratio test reduces to computing only two statistics

$$\ell_2 = \ell_{y_2} + \ell_{B_2} + \ell nP_2 - \ell nP_1 \tag{3.5a}$$

$$\ell_3 = \ell_{y_3} + \ell_{B_3} + \ell nP_3 - \ell nP_1 \tag{3.5b}$$

in which only ℓ_{y_2} and ℓ_{y_3} involve the raw data, ℓ_{B_2} and ℓ_{B_3} being fixed biases reflecting the average energy in the ensembles of unvoiced and voiced speech sounds. Letting

$$\lambda_u = -\ell_{B_2} - \ell nP_2 + \ell nP_1 \tag{3.6a}$$

$$\lambda_v = -\ell_{B_3} - \ell nP_3 + \ell nP_1 \tag{3.6b}$$

$$\lambda_{uv} = -\ell_{B_2} + \ell_{B_3} - \ell nP_2 + \ell nP_3 \tag{3.6c}$$

the classification rule reduces to the following:

(1) If: $\ell_{y_2} \leq \lambda_u$ and $\ell_{y_3} \leq \lambda_v$ (3.7a)
declare silence

(2) If: $\ell_{y_2} > \lambda_u$ or $\ell_{y_3} > \lambda_v$ and $\ell_{y_2} - \ell_{y_3} \geq \lambda_{uv}$ (3.7b)
declare unvoiced speech

(3) If: $\ell_{y_2} > \lambda_u$ or $\ell_{y_3} > \lambda_v$ and $\ell_{y_2} - \ell_{y_3} < \lambda_{uv}$ (3.7c)
declare voiced speech

In order to simplify the test further it is noted from (3.2) that the bias terms ℓ_{B_2} and ℓ_{B_3} are related to the energy in the ensemble of unvoiced speech and voiced speech sample functions. If a global average is taken, the voiced speech spectrum will have significantly more energy than that of unvoiced speech which would contribute a negative bias in favour of the unvoiced speech hypothesis. Using this bias would be valid if voiced speech were truly stationary. In fact however not only do the spectral properties change from frame to frame but more importantly the amplitude undergoes a slowly increasing and decreasing modulation at the beginning and ending of a voiced sound. Since 10-20 msec frames of speech represent the data base upon which a classification is to be made, then from a sample function point of view the energy in a frame of unvoiced speech or a frame of voiced speech could be comparable. The inclusion of the ensemble average energy bias term would therefore incorrectly favour unvoiced speech. Therefore the bias terms ℓ_{B_2} and ℓ_{B_3} must be assumed to be equal. Under this condition the thresholds reduce to

$$\lambda_u = -\ell_B - \ell nP_2 + \ell nP_1 \tag{3.8a}$$

$$\lambda_v = -\ell_B - \ell nP_3 + \ell nP_1 \tag{3.8b}$$

$$\lambda_{uv} = -\ell nP_2 + \ell nP_3 \qquad (3.8c)$$

where $\ell_B = \ell_{B_2} = \ell_{B_3}$ represents an unknown bias term related to the a priori knowledge of the energy in the unvoiced and voiced speech signals.

Although the Bayesian detection theory demands that the bias term and priori probabilities be calculated, a more practical method for determining the thresholds would be to train the system against noise and then choose those values that keep the false alarm rate at a value consistent with the system objectives. For example a much greater penalty is paid for failing to detect speech than falsely classifying noise as speech. Therefore the thresholds most likely should be set close to the 1-sigma values of ℓ_{y_2} and ℓ_{y_3} obtained during the noise training phase. This strategy is ideal for self-adaptive tracking of the noise statistics should they be non-stationary. The voicing threshold λ_{uv} is most reasonably approximated by zero when the signal-to-noise ratio is large or the noise is white. When this is not the case, this threshold can also be trained to the 1-sigma value of $\ell_{y_2} - \ell_{y_3}$.

As a result of the preceding analysis the only statistics that must be calculated at each frame time are the correlations

$$\ell_{y_i} = \sum_{n=1}^{N} y(n)\, \hat{s}_i(n) \qquad i=2,3 \qquad (3.9)$$

where $y(n)$ is the raw speech plus noise data and $\hat{s}_i(n)$ is the linear least squares unrealizable estimate of $s_i(n)$ given that hypothesis H_i is true. Since

the unvoiced and voiced speech waveforms are quasi-stationary the filter that results in $\hat{s}_i(n)$ given that $y(n) = s_i(n) + w(n)$ has the transfer function

$$H_i(f) = \frac{G_i(f)}{G_i(f) + N_o/2} \tag{3.10}$$

The filters defined by (3.10) obtain enhanced discrimination against noise by passing only those frequencies where the signal power is substantially larger than the noise power. Enhanced voiced-unvoiced discrimination depends on the implicit orthogonality of the two random processes as reflected by the degree to which the spectral densities are correlated. Both of these detection statistics can be improved by capitalizing on the quasi-periodic properties of voiced speech. If the voiced speech process is periodic with period T then the voiced speech power spectrum is more accurately represented by

$$G_3(f) = C(f;T)\ G_v(f) \tag{3.11}$$

where $G_v(f)$ represents the gross properties of the spectral envelope and $C(f;T)$ is a comb filter reflecting the fine structure of the periodic spectrum. If the period is maintained for M periods then

$$C(f;T) = \frac{1}{M}\ \frac{\sin(\pi Mf/F)}{(\pi f/F)}\ \cdot \exp\ [j\,\pi(M-1)\ f/F] \tag{3.12}$$

where F=1/T represents the pitch frequency. Not only does the comb filter enhance the voiced speech-to-noise ratio but it also increases the orthogonality of the voiced and unvoiced spectra. In order to exploit the additional discrimination implicit in the comb filter it is necessary that the pitch period be known. A discussion of how the pitch is to be determined will be deferred to a later section.

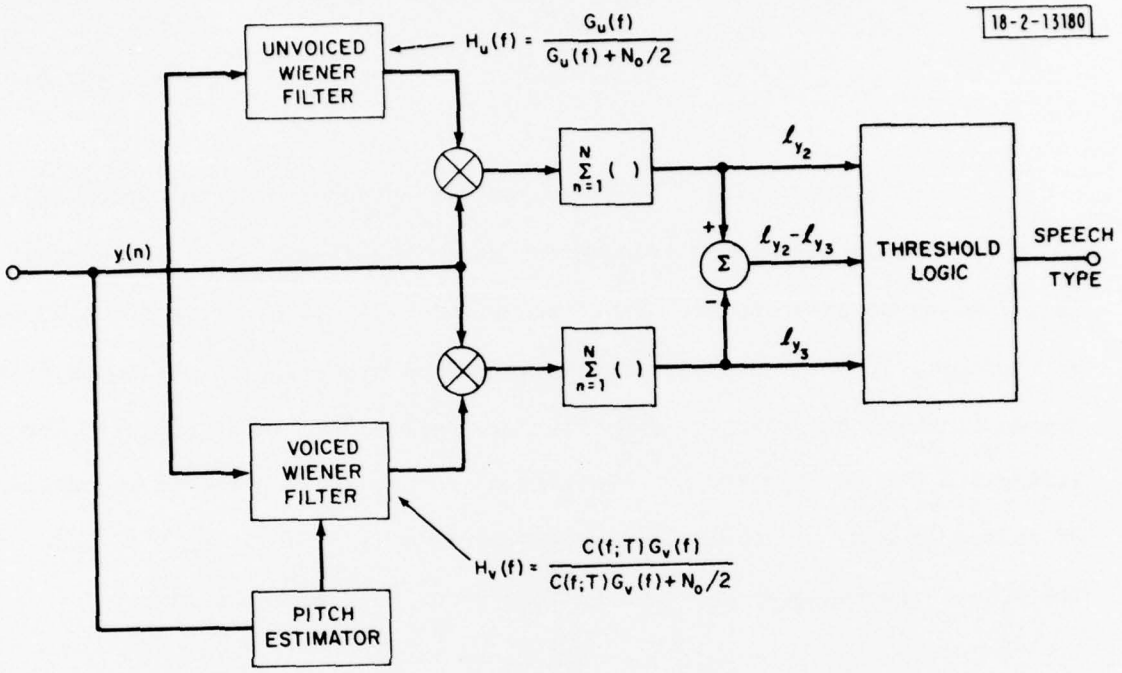


Fig. 1. The optimum speech classifier against white noise.

Subject to the assumptions that the envelopes of the unvoiced and voiced speech power spectra are known and that the pitch period for voiced speech can be estimated then the optimum classifier can be implemented as shown in Figure 1.

Of course all of this information is not available a priori and it will be necessary to introduce approximations to the filtering and estimation operations while maintaining the basic structure of the estimator-correlator receiver. This will be the goal of the next section.

IV. PRACTICAL IMPLEMENTATION OF THE ESTIMATOR-CORRELATOR SPEECH CLASSIFIER AGAINST WHITE NOISE

For voiced speech the optimum minimum mean squared error filter has the transfer function

$$H_V(f) = \frac{G_V(f)\, C(f;T)}{G_V(f)C(f;T) + N_o/2} \tag{4.1}$$

which passes those frequencies at which the signal power is substantially larger than the noise power and rejects all others. Certainly the comb filter in the denominator contributes to the definition of those frequencies at which noise rejection should occur. However, in white noise approximately the same rejection performance can be obtained by a cascade combination of the comb filter and the least squares filter designed on the basis of simply the spectral envelope. Therefore the voiced speech estimator filter is taken to be

$$H_V(f) = C(f;T) \cdot \frac{G_V(f)}{G_V(f) + N_o/2} \tag{4.2}$$

For unvoiced speech the estimator filter is

$$H_u(f) = \frac{G_u(f)}{G_u(f) + N_o/2} \tag{4.3}$$

Setting i=2 for unvoiced and i=3 for voiced, the Wiener filters based on the spectral envelopes for both cases can be written as

$$H_i(z) = \sum_{k=-\infty}^{\infty} a_k^i \, z^{-k} \tag{4.4}$$

where the coefficients a_k^i satisfy the Wiener-Hopf equation

$$\sum_{k=-\infty}^{\infty} a_k^i \, [R_i(k-j) + \sigma^2 \, \delta(k-j)] = R_i(j) \qquad -\infty < j < \infty \tag{4.5}$$

where $\sigma^2 = (N_o/2)F_s$ represents the energy in the noise process (F_s is the sampling rate) and where $R_2(k)$, $R_3(k)$ are the sampled data correlation functions corresponding to the power spectra $G_u(f)$, $G_v(f)$ respectively. In practice the correlation functions can be suitably truncated and then (4.4) can be efficiently solved using the Levinson recursion[9]. Of course the solution requires that the correlation functions for an ensemble of unvoiced and voiced speech sample functions be computed for a large class of utterances and a large class of speakers. In order to bootstrap the system initial classification would have to be done manually which would be extremely tedious and time consuming. In order to avoid this problem a more practical and robust strategy is proposed based on the well known global properties of unvoiced and voiced

speech spectra and a close examination of the filtering operation defined in (4.2) and (4.3).

The essence of the Wiener filter is to pass those frequencies at which the speech power is substantially larger than the noise power. As a good first approximation it seems reasonable to approximate the Wiener filter by a passband filter that passes "most" of the energy in an unvoiced or voiced speech sound. For unvoiced speech it can be assumed that "most" of the energy will be above 1000 Hz while for voiced speech "most" of the energy will be below 2000 Hz. While restricting the estimator filters to these frequencies improves the detection SNR of unvoiced and voiced speech, of at least equal importance is the ability of the unvoiced filter to reject voiced speech and vice versa. Since the first formant of voiced speech is approximately 1000 Hz then if the cutoff of the unvoiced speech filter is above 1250 Hz then most of the unvoiced speech energy will pass through the filter while a large fraction of a voiced speech signal will be attenuated. Similarly if the cutoff of the voiced speech signal is above 2000 Hz then most of its energy will pass through the voiced filter while a substantial fraction of an unvoiced speech signal will be attenuated. From this point of view it can be seen that it is crucial that the input data to the classifier not be preemphasized since the higher formants of a voiced speech signal would take on the attributes of an unvoiced speech waveform at the expense of good classifier performance. Therefore if pre-emphasis is to be used for speech analysis and synthesis the data will have to undergo digital deemphasis prior to speech classification.

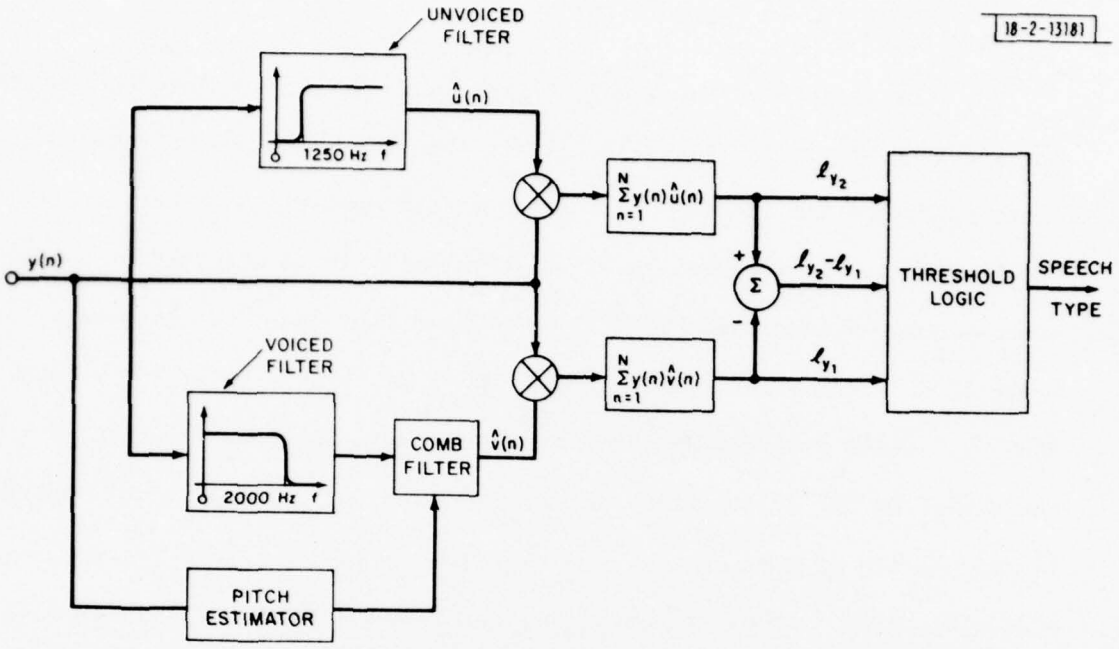


Fig. 2. Practical realization of the optimum speech classifier.

On the basis of the preceding arguments the Wiener filter for unvoiced speech will be approximated by a high pass linear phase digital filter whose cutoff frequency is below 1250 Hz. For voiced speech a lowpass linear phase digital filter having a cutoff frequency above 2000 Hz will be used. The linear phase requirement is essential since the temporal properties of the waveforms must be preserved in order that a meaningful correlation operation be obtained. The practical implementation of the optimum classifier against white noise is shown in Figure 2. The detailed characteristics of the linear phase filters are provided in the Appendix.

Implicit in the realization illustrated in Figure 2 is the estimation of the pitch period of a voiced waveform so that the additional discrimination inherent in the comb filter can be exploited. A further simplification in processor complexity can be obtained simply by omitting the comb filter and relying on the spectral orthogonality of the two speech types. However, since the periodicity of the voiced speech process is a potentially powerful classification discriminant, for theoretical completeness, it is worthwhile to develop a practical algorithm to exploit it. Since this necessitates an estimate of the pitch period, a brief exposition of an optimum pitch estimation algorithm will be presented.

V. OPTIMUM PITCH ESTIMATION

Voiced speech was modelled as a periodic random process in the sense that $R_v(k) = R_v(k+T)$ for some pitch period T. This means that almost every sample function in the ensemble is periodic with period T. Therefore the voiced speech signal, v(n), can be modelled as

$$v(n) = q(n)_{\text{mod}T} \tag{5.1}$$

where q(1), q(2),..,q(T) are completely unknown. Of course to be faithful to the random process formulation of voiced speech, the quantities q(k) should be treated as correlated random variables. However to keep the estimation problem mathematically tractable the correlation properties will be ignored at first. The voiced speech data are therefore taken to be

$$y(n) = v(n) + w(n) \tag{5.2}$$

where w(n) represents white Gaussian noise and v(n) is given by (5.1). Based on N samples of this data the parameters q(1), q(2),...,q(T) and T are to be estimated.

The above formulation of the pitch estimation problem was formulated and solved by Wise, Caprio and Parks[10]. Using the maximum likelihood estimation rule they minimized the cost function

$$D(\underline{q},T) = \sum_{n=1}^{N} [y(n) - v(n)]^2$$

$$= \sum_{n=1}^{N} y^2(n) - 2 \sum_{k=1}^{T} \sum_{m=0}^{M-1} y(k+mT)\, v(k+mT)$$

$$+ \sum_{k=1}^{T} \sum_{m=0}^{M-1} v^2(k+mT) \tag{5.3}$$

In order to simplify the derivation, it has been assumed that N=MT, M an integer*. From the periodicity condition $v(k+mT) = q(k)_{\text{mod}T}$, then (5.3) reduces to

*The more general case is tedious and contributes little to the final result.

$$D(\underline{q},T) = \sum_{n=1}^{N} y^2(n) - 2 \sum_{k=1}^{T} q(k) \sum_{m=0}^{M-1} y(k+mT) + M \sum_{k=1}^{T} q^2(k) \qquad (5.4)$$

Since the basic voiced speech waveform $q(1),\ldots,q(T)$ has been assumed completely unknown (i.e., the correlation properties have been ignored *) then, for the fixed T, the minimizing values are obviously

$$\hat{q}(k) = \frac{1}{M} \sum_{m=0}^{M-1} y(k+mT) \qquad (5.5)$$

The estimate of the voiced speech waveform is therefore

$$\hat{v}(n|N) = \hat{q}(k)_{\text{mod}T} \qquad (5.6)$$

where the notation $\hat{v}(n|N)$ is used to denote the fact that all N measurements $y(1), y(2),\ldots,y(N)$ are used in developing the estimate of the voiced speech waveform $v(n)$, $n \leq N$. In that sense, the estimator is unrealizable†. The corresponding minimum value of the likelihood function is

$$D(T) = \sum_{n=1}^{N} [y(n) - \hat{v}(n|N)]^2 \qquad (5.7a)$$

$$= \sum_{n=1}^{N} y^2(n) - \sum_{n=1}^{N} \hat{v}^2(n|N) \qquad (5.7b)$$

Since $\hat{v}(n|N)$ can be interpreted as the output of a comb filter tuned to pitch period T when $y(n)$ is the input, then the second term in (5.7b) simply

*The more general case is treated by McAulay[11].

† A realizable estimator that uses only the data up to time n is

$$\hat{v}(n|n) = \frac{1}{M} \sum_{m=0}^{M-1} y(n-MT).$$

represents the energy at the output of this comb filter. Therefore the opitmum estimate of the pitch period can be obtained by constructing a bank of comb filters each tuned to a slightly different pitch period and choosing as the estimate the pitch corresponding to the comb filter for which the output energy is largest.

It is important to keep in mind the fact that voiced speech signals are at best quasi-periodic; hence, there is a definite limitation on the number of periods over which the averaging process is a meaningful operation. Since values of the pitch frequency generally fall within the range 70-300 Hz corresponding to pitch periods 3-15 ms long, and since the time required for a significant alteration in the vocal tract is approximately 20 ms, there can be 1-7 repetitions of the voiced speech waveform. Therefore the number of periods over which the data is averaged is a design parameter that must be chosen to carefully trade off the estimation accuracy and the quasi-periodic nature of the voiced speech waveform.

A particularly important practical case corresponds to the assumption that the voiced speech waveform is periodic for two successive periods. In this case from (5.5) and (5.6) the maximum likelihood estimate of the voiced speech signal is

$$\hat{v}(n|N) = \frac{1}{2}\,[y(n) + y(n-T)] \tag{5.8}$$

which from (5.7a) results in the residual error

$$D(T) = \sum_{n=1}^{N} [y(n) - \hat{v}(n|N)]^2 = \frac{1}{4} \sum_{n=1}^{N} [y(n) - y(n-T)]^2 \tag{5.9}$$

The estimate of the pitch period is then the value of T that minimizes D(T). This criterion has already been proposed for pitch estimation by Moorer[12] and Ross et al.[13] except that the squared difference has been approximated by the absolute magnitude difference function in order to achieve greater dynamic range and computational speed. Experimental results have shown that the quality of the pitch estimates is roughly equivalent to that of the cepstral method and successful operation has also been demonstrated in strong noise environments. For this reason it is conjectured that the (5.5)-(5.7) represent a possible solution to the problem of robust pitch estimation. To see this suppose that the true pitch period is T_o. Then the observed data is

$$y(n) = v(n;T_o) + w(n) \tag{5.10}$$

where $v(n;T_o) = q(k)_{\text{mod}T_o}$. The output of the comb filter tuned to pitch period T is

$$\hat{v}(n;T) = \frac{1}{M} \sum_{m=0}^{M-1} v(n-mT;T_o) + \frac{1}{M} \sum_{m=0}^{M-1} w(n-mT) \tag{5.11}$$

The noise signal at the output of the comb filter is

$$\eta(n;T) = \frac{1}{M} \sum_{m=0}^{M-1} w(n-mT) \tag{5.12}$$

As long as the correlation time of the noise process is less than the minimum pitch period of interest, then if w(n) has variance σ^2, $\eta(n;T)$ will have variance σ^2/M. For the comb filter tuned to pitch T_o the output signal is

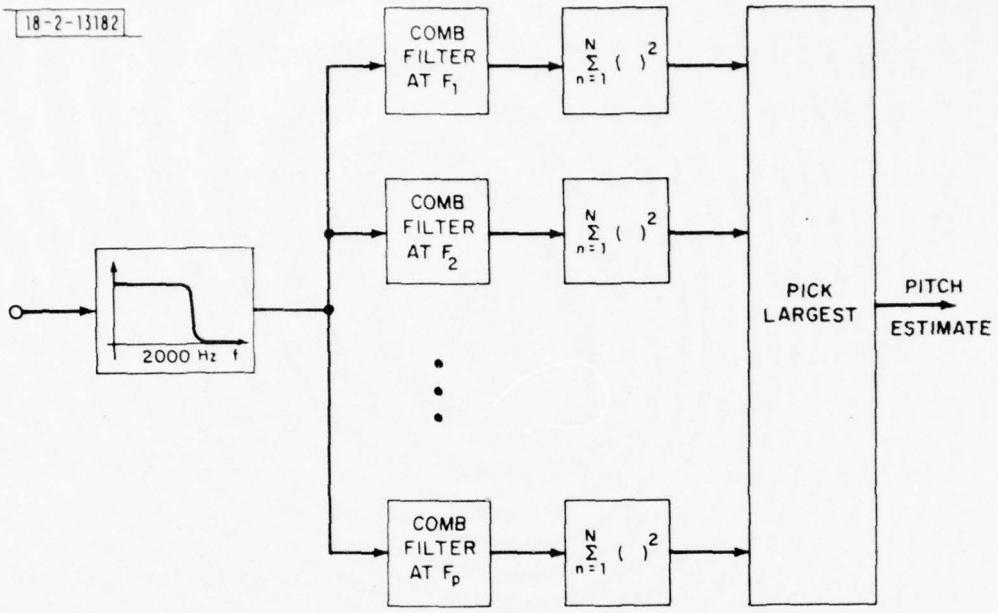


Fig. 3. Practical implementation of the optimum pitch estimator.

$$\hat{v}(n;T_o) = q(k)_{\text{mod}T_o} + \eta(n;T_o) \qquad (5.13)$$

Therefore there is an M:1 increase in signal-to-noise ratio as a result of using the comb filter. Applied to the two-pulse canceller in (5.10) (i.e., the AMDF) a 3 dB improvement in SNR is obtained for the class of noise processes whose correlation times are less than the minimum pitch period of interest.

Although originally proposed as a pitch estimation criterion based on ad hoc considerations, the maximum likelihood theory shows that the average squared difference function is optimum and robust when the voiced speech waveform is modelled as a deterministic quasi-periodic waveform with periodicity extending over two periods. The major limitation in using the two-pulse comb filter (i.e., the AMDF) is the not infrequent occurrence of pitch doubling which occurs when the voiced speech is periodic for at least four pitch periods. At the expense of increasing the length of the speech buffer, an M-pulse comb filter, $M \geq 3$, can be used to reduce the rate at which pitch doubling errors occur.

A further enhancement in the pitch estimate can be obtained by using the low pass voiced speech filter to increase the pitch estimator SNR. This corresponds to exploitation of the global correlation properties of voiced speech. The approximate matched filter configuration of the pitch detector is shown in Figure 3.

VI. THE OPTIMUM CLASSIFIER AGAINST COLOURED NOISE

There are several examples in which speech in non-white acoustic background noise can be effectively classified using the algorithm that was

defined to be optimum against white noise. In particular whenever the signal-to-noise ratio is high, the white noise classifier will yield acceptable performance. There are some cases, particularly if the SNR is low and the noise is highly correlated, where significant improvements can be achieved by taking the spectral characteristics of the noise into account. In this section the structure of the optimum classifier will be derived for the coloured noise case and then reasonable practical approximations will be deduced in order to simplify the complexity of the signal processor.

For this classification problem the data corresponding to hypothesis H_i is

$$y(n) = s_i(n) + w_c(n) + w(n) \qquad i=1,2,3 \qquad (6.1)$$

where $w_c(n)$ denotes the coloured noise present on all three hypotheses. Note that a white noise component, w(n), is also incorporated into the model to avoid mathematical problems relating to singular solutions. The standard approach to this problem is to precede all of the processing by a whitening filter and then apply the white noise solution. This was the approach taken by McAulay[11]. Although mathematically correct, this approach encounters practical difficulties because the whitening filter essentially preemphasizes the speech data. As has already been discussed, this can cause the higher formants of voiced speech to acquire the same attributes as unvoiced speech which makes classification difficult. McAulay and Yates[14] have derived an estimator-correlator classifier that does not require a whitening pre-filter. Drawing on their results and those developed in Section III two sufficient statistics are computed. They are

$$\ell_{zi} = \sum_{n=1}^{N} z(n)\ \hat{s}_i(n) \qquad i=2,3 \tag{6.2}$$

where

$$\hat{s}_i(n) = \sum_{k=-\infty}^{\infty} h_i(n-k)\ y(k) \quad i=2,3 \tag{6.3}$$

is the linear least squared error unrealizable estimate of $s_i(n)$ based on the data $y(n) = s_i(n) + w_c(n) + w(n)$ and where

$$z(n) = \sum_{k=-\infty}^{\infty} h_c(n-k)\ y(k) \tag{6.4}$$

is the result of passing $y(n)$ through the clutter rejection filter $h_c(n)$. It has been implicitly assumed that the speech and noise processes are independent and quasi-stationary. The transfer functions of the filters are[14]

$$H_i(f) = \frac{G_i(f)}{G_i(f) + G_c(f) + N_o/2} \qquad i=2,3 \tag{6.5}$$

$$H_c(f) = 1 - \frac{G_c(f)}{G_c(f) + N_o/2} = \frac{N_o/2}{G_c(f)+N_o/2} \tag{6.6}$$

where $G_c(f)$, $G_2(f)$, $G_3(f)$ represent the power spectra for the coloured noise, unvoiced speech and voiced speech processes respectively. The second term in (6.6) is precisely the linear least squares unrealizable estimator of $w_c(n)$ based on the signal $w_c(n) + w(n)$. Therefore the clutter filter attempts to remove the coloured noise from the data before performing the correlation operation. The optimum classifier structure is shown in Figure 4. The

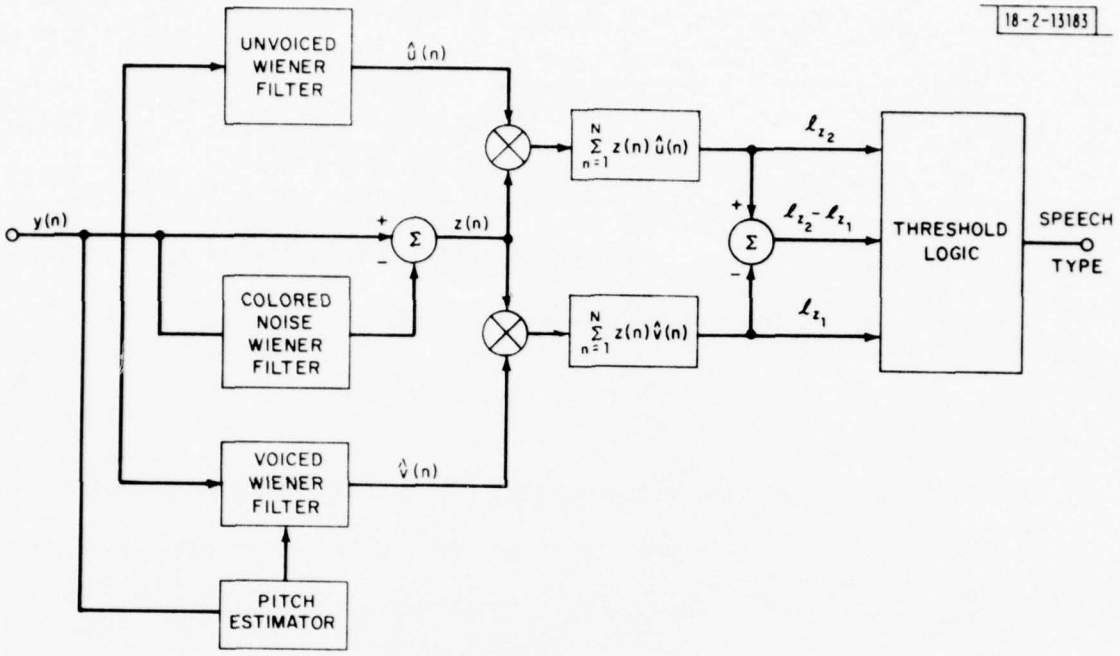


Fig. 4. The optimum speech classifier against coloured noise.

classification rule is similar to that derived for white noise, equation (3.7) except that the sufficient statistics are now ℓ_{z_2} and ℓ_{z_3} instead of ℓ_{y_2} and ℓ_{y_3}.

VII. PRACTICAL IMPLEMENTATION OF THE ESTIMATOR-CORRELATOR SPEECH CLASSIFIER AGAINST COLOURED NOISE

The arguments for simplifying the processing of voiced and unvoiced speech proceed along the same lines as those made for the white noise case. In particular, if knowledge of pitch is available the spectral harmonics of voiced speech are matched by using a comb filter in cascade with the Wiener filter designed on the basis of the spectral envelope. Therefore the voiced speech estimator filter is

$$H_v(f) = C(f;\hat{T}) \frac{G_v(f)}{G_v(f) + G_c(f) + N_o/2} \tag{7.1}$$

where $C(f;T)$ is the comb filter tuned to the most recent pitch estimate, T. For unvoiced speech the estimator filter is*

$$H_u(f) = \frac{G_u(f)}{G_u(f) + G_c(f)} \tag{7.2}$$

Lacking knowledge of the exact form of $G_v(f)$ and $G_u(f)$ a good first approximation is to use the linear phase low pass (cutoff above 2000 Hz) and high pass (cutoff below 1250 Hz) filters in the voiced and unvoiced speech channels as was done in the white noise case. This insures the spectral orthogonality of the two speech channels and enhances the speech-to-noise ratio whenever the noise spectrum lies outside the filter passbands. For coloured noise, however, it is possible that all of the noise energy will lie within the filter passbands in which case no speech enhancement will occur if only the

*The effects of the artificial white noise term have been neglected at this point since there is no problem with singular solutions.

fixed filters are used. Somehow additional processing tuned to reject the clutter will have to precede the fixed filters in the speech channels. To develop a clue as to the form of the clutter processor it is necessary to reexamine (7.1) and (7.2). Letting $G_2(f) = G_u(f)$ and $G_3(f) = G_v(f)$ then the unvoiced and voiced speech Wiener filters can be written as

$$H_i(f) = \frac{G_i(f)}{G_i(f) + G_c(f)} \tag{7.3}$$

Realization of these filters requires that the speech and noise spectra be known. Since the noise statistics can be measured during the silent intervals it is reasonable to assume that the clutter spectrum is known. Unfortunately a priori estimates of the speech spectra are not available unless long term averages are determined from training sets. When detailed knowledge of the frequency distribution of the speech is unavailable a conservative approach is to model the speech as white noise thereby having a flat spectrum. Letting

$$G_i(f) = \alpha_i \qquad i = 2, 3 \tag{7.4}$$

and substituting this into (7.3) results in the filters

$$H_i(f) = \frac{\alpha_i}{G_c(f) + \alpha_i} \qquad i = 2, 3 \tag{7.5}$$

Since $H_i(f) \approx 0$ whenever $G_c(f) >> \alpha_i$ and $H_i(f) \approx 1$ whenever $G_c(f) << \alpha_i$, (7.5) can be interpreted as a notch filter tuned to reject "most" of the clutter energy. When the speech-to-noise ratio (SNR) is large little

clutter rejection is needed and α_i should be large since this results in a passband filter. When the SNR is small, then the clutter must be rejected whatever the cost in speech distortion which necessitates a small value for α_i. It follows, therefore, the parameter α_i should be proportional to the speech-to-noise ratio. Since the clutter power is known from the silent intervals, estimates of the SNR can be made from the data frame being analyzed. In this mode the distinction between voiced and unvoiced speech disappears and only a single parameter value and clutter filter need be determined. In this sense the clutter filter represents an adaptive prefilter whose output, in a conservative sense, represents the best available estimate of the speech waveform.

The results of this discussion are summarized in Figure 5 which shows the practical realization of the optimum classifier operating against a coloured noise background. Except for the clutter filters in the reference and speech channels the processing is identical to that used in the white noise case. Since selection of the tuning parameters α_c and α_s depends on the noise statistics further discussion regarding their selection will be deferred to the section on experimental results.

The only problem that remains to be discussed is the calculation of the clutter filter impulse response from (7.5). The most straightforward approach is to solve the Wiener-Hopf equation

$$\sum_{k=-\infty}^{\infty} a_k [R_c(k-j) + \alpha\delta(k-j)] = \alpha\delta(j) \qquad -\infty < j < \infty \qquad (7.6)$$

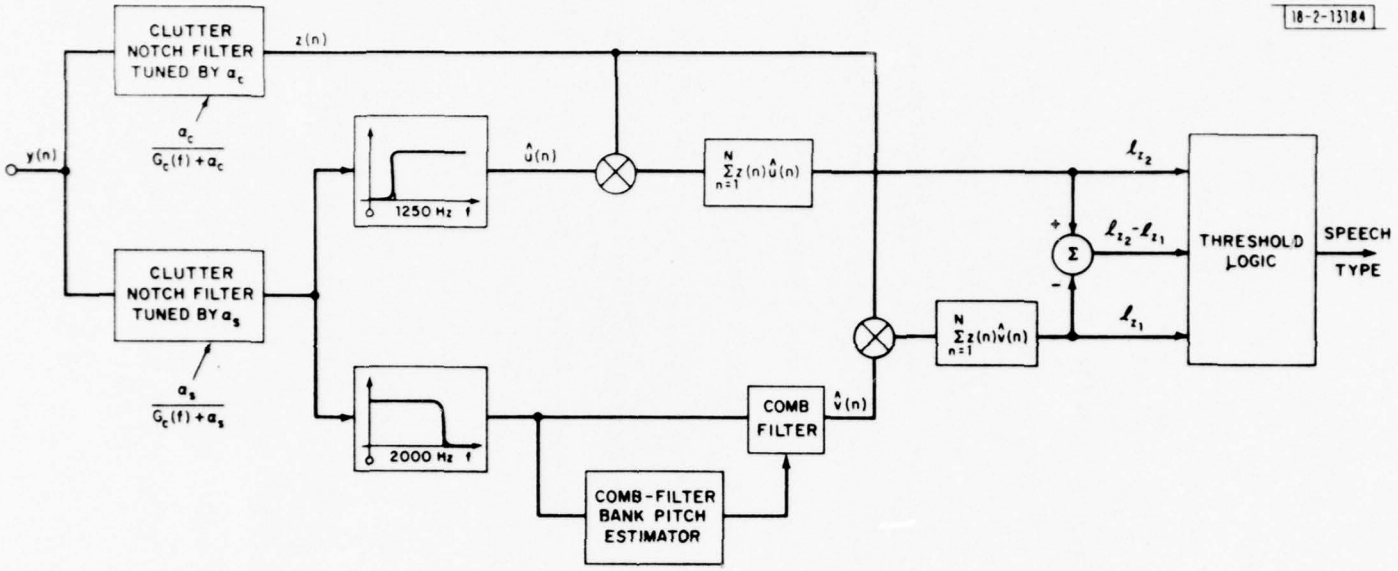


Fig. 5. Practical realization of the optimum speech classifier against colou red noise.

If the impulse response is truncated at $\pm p$, the 2p+1 coefficients, a_k, can be found by solving (7.6) numerically using the Levinson recursion. Another approach is to fit an all pole spectrum to $G_c(f) + \alpha$ using Linear Prediction techniques and use the spectral coefficients to determine the clutter filter. For this method the LPC spectral estimate of $G_c(f) + \alpha$ can be obtained by solving

$$\sum_{k=1}^{p} a_k \left[R_c(k-j) + \alpha\delta(k-j)\right] = R_c(j) \qquad 1 \leq j \leq p \qquad (7.7)$$

This equation can be solved efficiently using the Levinson Recursion and results in a p-pole fit to the clutter spectrum. The estimated spectrum is

$$\widehat{G_c(z) + \alpha} = \frac{\sigma}{A(z)A^*(z)} \qquad (7.8)$$

where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (7.9)$$

which corresonds to the Inverse Filter in the usual LPC analysis. Substituting (7.8) into (7.5) results in the Wiener filter

$$H(z) = \frac{\alpha}{\sigma} A(z) A^*(z) \qquad (7.10)$$

Letting y(n) denote the input sequence and $\hat{s}(n)$ the output sequence then

$$\hat{S}(z) = \frac{\alpha}{\sigma} A(z) A^*(z) Y(z)$$

$$= \frac{\alpha}{\sigma} A(z) X(z) \qquad (7.11)$$

where

$$X(z) = A^*(z)\, Y(z) \tag{7.12}$$

Since the LPC coefficients $\{a_k\}$ are real

$$A^*(z) = 1 - \sum_{k=1}^{p} a_k z^k \tag{7.13}$$

and

$$x(n) = y(n) - \sum_{k=1}^{p} a_k y(n+k) \tag{7.14}$$

$$\hat{s}(n) = \frac{\alpha}{\sigma} \left[x(n) - \sum_{k=1}^{p} a_k x(n-k)\right] \tag{7.15}$$

Therefore the unrealizable Wiener filter can be implemented by the cascade combination of an inverse filter that operates on p samples of future data and an inverse filter that operates on p samples of past data. Therefore a p-sample buffer must be available to provide for the future data. The advantage of this approach is that the length of the impulse response is completely determined on the basis of the number of poles required to fit the clutter spectrum.

VIII. EXPERIMENTAL RESULTS

The signal processing concepts developed in the previous sections were evaluated experimentally using speech data that was corrupted by Airborne Command Post (ACP) noise. Not only does this provide a good pedagogical tool for illustrating the filtering ideas but it represents an important real-world speech encoding environment which is not adequately solved using state-of-the-art vocoder technology.

The noisy speech data was sampled every 132 μsec (7575 Hz) and 158 samples were collected to define a 20 millisec. frame. Figure 6a illustrates a 20 millisec sample function of ACP noise. Figure 6f is a plot of the magnitude of its Fourier Transform measured in dB. The correlation function of the mth frame (i.e., the current frame) of noise data was computed from

$$R_y(k;m) = \sum_{n=0}^{N-1-k} x(n)\, x(n+k) \qquad k=0,1,\ldots,p;\ m=1,2,\ldots \qquad (8.1)$$

where x(n) is the Hamming weighted version of the input data y(n). A first order smoothed correlation function was then computed from

$$\bar{R}_c(k;m) = \frac{1-\gamma}{1-\gamma^m}\left[R_y(k;m) + \gamma\, R_c(k;m-1)\right] \qquad (8.2)$$

In general the weighting constant γ should be chosen to reflect the quasi-stationarity of the noise random process. For ACP noise $\gamma = .95$ was chosen arbitrarily and seemed to produce good results.

From (6.6) the clutter filter in the reference channel was given by

$$H_c(z;m) = \frac{\alpha_c(m)}{G_c(z)+\alpha_c(m)} \qquad (8.3)$$

The impulse response was found using Linear Prediction techniques as described in the previous section. This necessitates solving the Wiener-

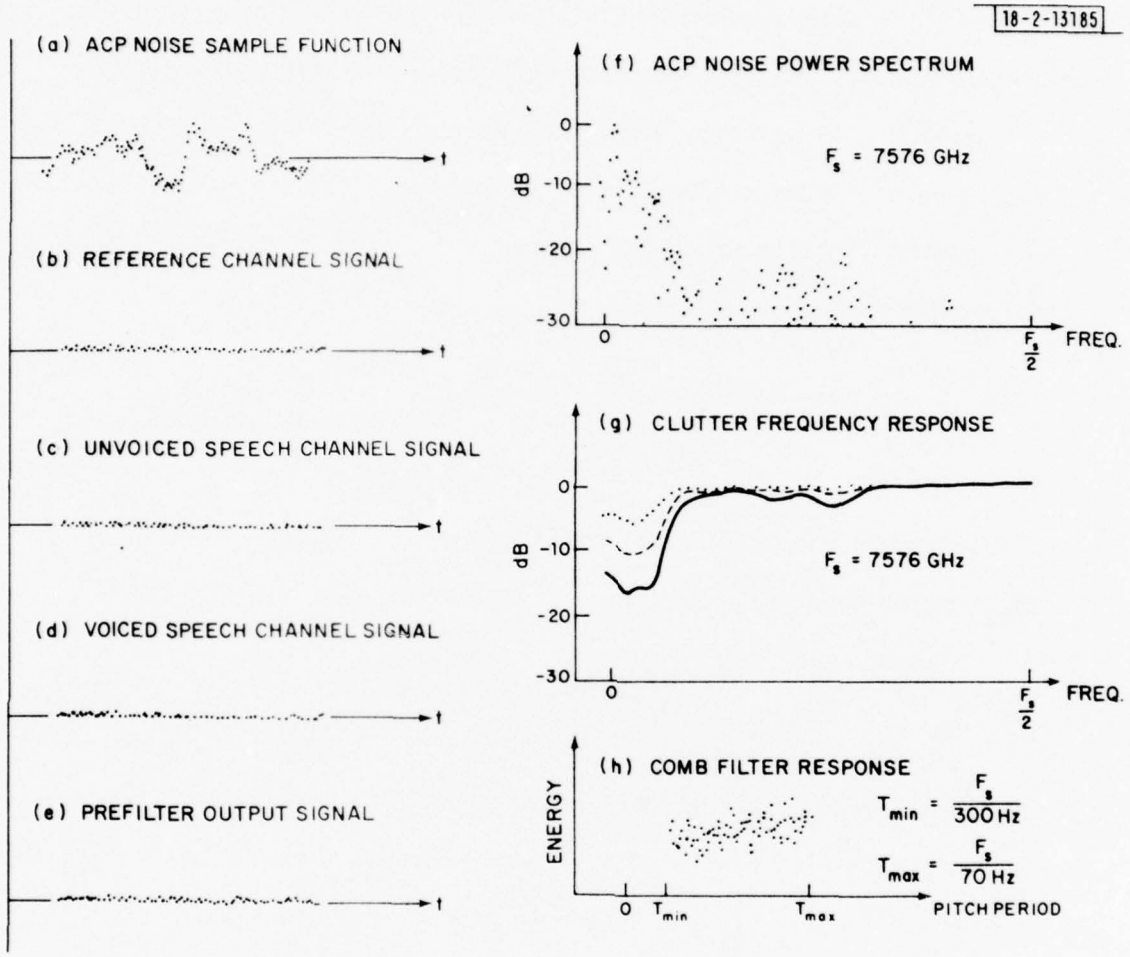


Fig. 6. Processor response due to noise.

Hopf prediction equation

$$\sum_{k=1}^{p} a_k \left[\bar{R}_c(k-j;m) + \alpha_c(m)\ \delta(k-j)\right] = \bar{R}_c(j;m) \quad 1 \leq j \leq p \qquad (8.4)$$

using the long term averaged correlation function computed at the last frame (i.e., the mth frame). A whole class of clutter filters can be obtained simply by varying the parameter $\alpha_c(m)$. Typical transfer functions from this class are shown in Figure 6g for three values of α_c. It was found that the clutter filter defined for the value $\alpha_c(m) = \bar{R}_c(\emptyset;m)$ worked well for ACP noise. For other noise types other values would probably be more appropriate. A little experimentation is therefore required to tune the clutter filter to different noise processes.

The unvoiced and voiced speech channels are preceded by another clutter rejection filter given by (7.5), namely

$$H_s(z;m) = \frac{\alpha_s(m)}{G_c(z) + \alpha_s(m)} \qquad (8.5)$$

where α_s is chosen to be proportional to the speech-to-noise ratio measured for the current frame of data (i.e., the mth frame). Since $R_y(\emptyset;m)$ represents a measure of the speech plus noise energy for the current frame of data and since $\bar{R}_c(\emptyset;m)$ represents a measure of the long term averaged noise energy, then a reasonable estimate for the speech-to-noise energy is

$$\hat{\xi}(m) = R_y(\emptyset;m) - \bar{R}_c(\emptyset;m) \qquad (8.6)$$

It is possible that the energy in any one 20 millisec sample function will be less than the average clutter energy, especially if that sample function contains noise alone or noise plus unvoiced speech. Therefore provision must be made to bound the clutter notch parameter α_s away from zero. A reasonable scheme is to pick

$$\alpha_s(m) = \max [\hat{\xi}(m), \alpha_c(m)] \tag{8.7}$$

which guarantees that the speech-clutter-filter notch will never be deeper than that in the reference channel. As before the impulse response was found using the Linear Prediction power spectrum which was obtained by solving the Wiener-Hopf predictor equation (8.4) using α_s instead of α_c.

The output of the speech clutter filter was then used as the input to the high-pass and low-pass filters characterizing the unvoiced and voiced speech processing channels respectively. The filters were both 21-tap linear phase digital filters designed using the Parks-McClellan algorithm[15]. The impulse responses and frequency characteristics are specified in the Appendix. No attempt was made to optimize the filter design. The outputs of the reference channel clutter filter $z(n)$ and the unvoiced and voiced speech filters $\hat{u}(n)$, $\hat{v}(n)$ are shown in Figures 6b, 6c, 6d. According to equation (6.2) the outputs of the speech filters were then correlated with the output of the reference channel clutter filter to form the detection statistics:

$$\ell_1(m) = \sum_{n=1}^{N} z(n)\hat{u}(n) \tag{8.8a}$$

$$\ell_2(m) = \sum_{n=1}^{N} z(n)\hat{v}(n) \tag{8.8b}$$

$$\ell_3(m) = \ell_u(m) - \ell_v(m) \tag{8.8c}$$

It should be noted that the comb filter has been left out of the voiced speech processing channel. This decision was made to show that good classifier performance could be obtained without having to make a pitch estimate which simplifies the classifier processing which is necessary for some applications.

The detection thresholds were obtained by driving the system with ACP noise for 15 data frames (.3 sec). This is the only training cycle required by the processor and should be relatively easy to meet in practice because there is always a speech free interval before a talker actually speaks into the encoding device after having turned the machine on. Averaged detection statistics for the training noise are computed from

$$\bar{\ell}_i(m) = \frac{1-\gamma}{1-\gamma^m}\left[\ell_i(m) + \gamma\bar{\ell}_i(m-1)\right] \qquad i=1,2,3 \tag{8.9}$$

with $\gamma = .95$ as before. The detection thresholds were then chosen to be

$$\begin{aligned} \lambda_i(m) &= 1.5\ \bar{\ell}_i(m) \qquad i=1,2 \\ \lambda_3(m) &= \bar{\ell}_1(m) - \bar{\ell}_2(m) \end{aligned} \tag{8.10}$$

which allows for moderate statistical fluctuations. After the first 15 data frames of noise have been processed (m=15) and the initial threshold setting computed, the classification process is initiated.

The next frame of data is processed and the detection statistics $\ell_i(m+1)$ are computed. If $\ell_1(m+1) < \lambda_1(m)$ and $\ell_2(m+1) < \lambda_2(m)$ then the data are classified as silence and the clutter correlation function, (8.2) and the detection thresholds (8.9) and (8.10) are up-dated. If $\ell_1(m+1) > \lambda_1(m)$ or $\ell_2(m+1) > \lambda_2(m)$ then speech is declared present and neither the clutter correlation function nor the detection thresholds are changed. No up-dating is done until the next frame of silence is detected. This procedure allows the classifier to track noise processes whose statistics vary slowly with time. Such a classifier structure is often referred to as a decision-directed detector since it tells itself when to alter its structure. It becomes evident therefore that the detection thresholds should be set low even at the expense of a high false alarm rate (declaring noise as speech is a false alarm). It would be a more serious error if the classifier declared speech as noise since then all the clutter filters and detection thresholds would be tuned to reject speech. Fortunately this malign event rarely occurred for ACP noise and when it did the noise always completely overpowered the speech so that little change in the filter structures occurred.

The effects of the three filtering channels on the three speech types will be examined for some typical cases to develop a feeling for the classifier operation. Figure 6a is a plot of a 20 millisec input sample function of ACP noise. Figure 6f is the corresponding short-term power spectrum. Figure 6g is a plot of the adaptive clutter filter transfer function in the reference channel (the adaptive prefilter).

For ACP noise input, it has adapted in such a way as to make a -10dB null at the clutter frequencies. Figures 6b, 6c and 6d show the respective outputs of the reference channel, the high-pass filtered unvoiced speech channel and the low-pass filtered voiced speech channel.

As was described in the previous section the output of the speech channel clutter filter represents a minimum mean squared error estimate of the input speech. Figure 6e shows a plot of the prefilter output in response to ACP noise at the input. Of course, with high probability the classifier will classify the frame as silence, hence one has the option of setting the prefilter output to zero which removes the residual noise completely.

Although the comb filter discriminator was not used in the classifier it remains of interest to evaluate the robustness of the maximum likelihood pitch estimator in ACP noise. This was done by applying the output of the low pass filter, $\hat{v}(n)$, to a bank of two-pulse comb filters covering the range from 70 to 300 Hz. Figure 6h is a plot of the energy at the output of the comb filters as a function of the pitch period for the ACP noise sample.

The same sequence of data are plotted in Figures 8 and 9 for 20 millisec frames of unvoiced and voiced speech respectively. Figures 7a and 7f show that the unvoiced speech-to-noise ratio is less than 0dB (it is roughly -3dB) yet Figure 7e shows that the prefilter has removed a significant portion of the clutter waveform while allowing the unvoiced speech waveform to pass relatively undisturbed. Figures 8a

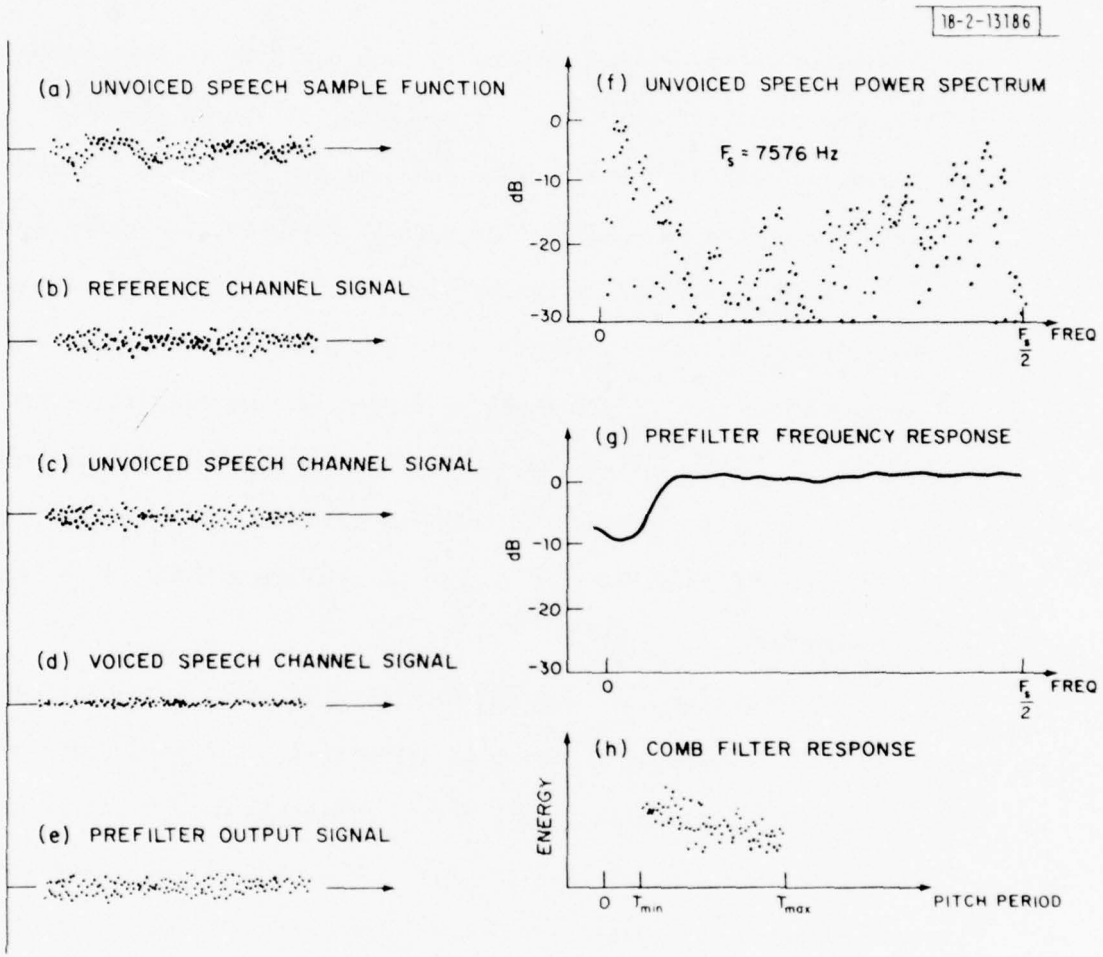


Fig. 7. Processor response due to unvoiced speech.

and 8f show that the voiced speech-to-noise ratio is quite large (it is roughly 9dB). Figure 8g shows that the prefilter transfer function is adjusted to allow most of the speech to pass even though its spectrum overlaps that of the ACP noise. This shows the advantage of the adaptive prefilter. Had a fixed clutter filter been used, the voiced speech waveform would have been distorted unnecessarily. Figure 8h shows that the pitch estimate is perturbed very little by the presence of ACP noise. In general it was found that the only significant pitch errors were the effects of pitch doubling which occurred intermittently near the ends of a voiced sound. Figure 8e shows how the prefilter attempts to reproduce the voiced speech waveform.

Having established the basic characteristics of the classifier the next step is to evaluate the frame-to-frame performance when an ACP noise corrupted utterance is applied to the input. Classification errors were obtained by determining the true speech type by visually examining the waveform, power spectrum and comb filter energy contour for each 20 millisec sample function. Statistics were accumulated for a total of 3 utterances spoken by 3 male speakers in different ACP noise environments. The results are tabulated in Table 1. From these results the false alarm probability (declare speech given silence) is estimated to be 9.4%. The miss probability (declare silence given speech) is 2.3%. The misses mainly occurred for unvoiced speech that had been completely overpowered by the noise ($\sim$ -10dB SNR). Erroneous classifications (voiced$\leftrightarrow$unvoiced) occurred at the rate of 3%. Whenever

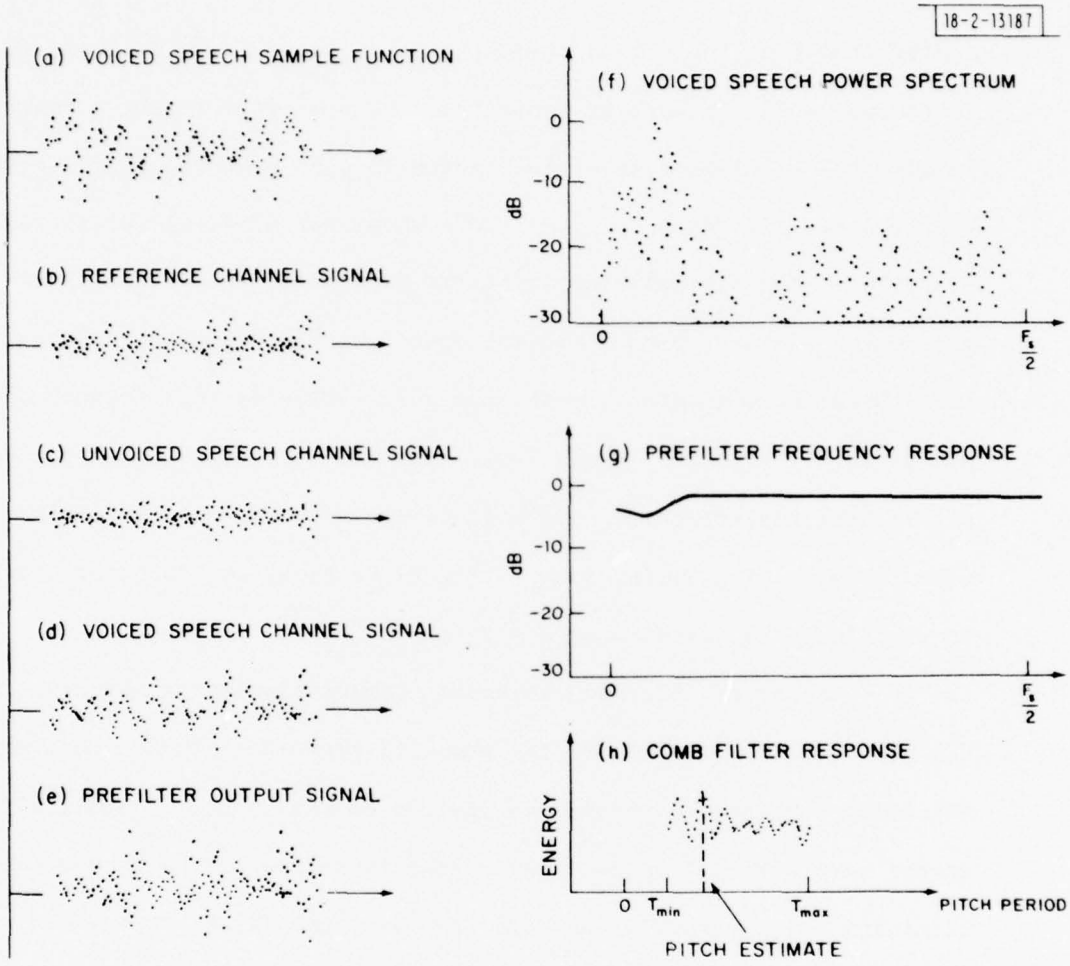


Fig. 8. Processor response due to voiced speech.

TABLE 1

CLASSIFIER PERFORMANCE STATISTICS

TRUE \ ESTIMATED	SILENCE	UNVOICED	VOICED
SILENCE	405	14	24
UNVOICED	4	43	2
VOICED	1	5	170
UNVOICED-VOICED	0	0	6

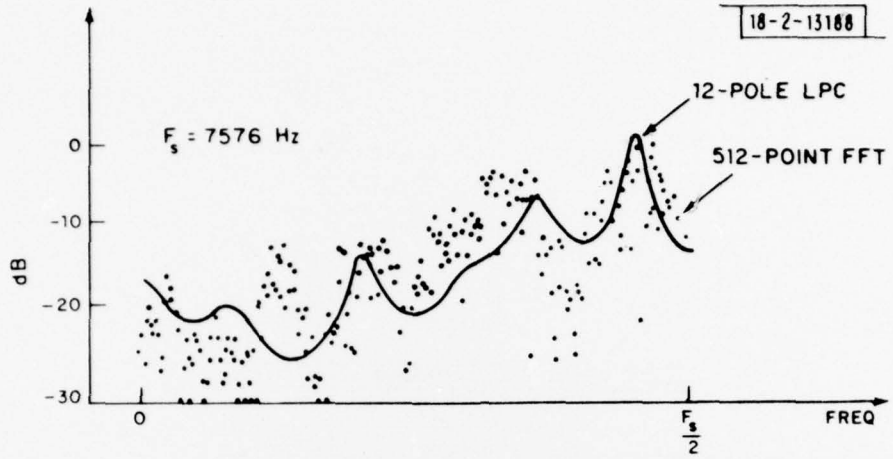


Fig. 9. Unvoiced speech prefilter output spectra.

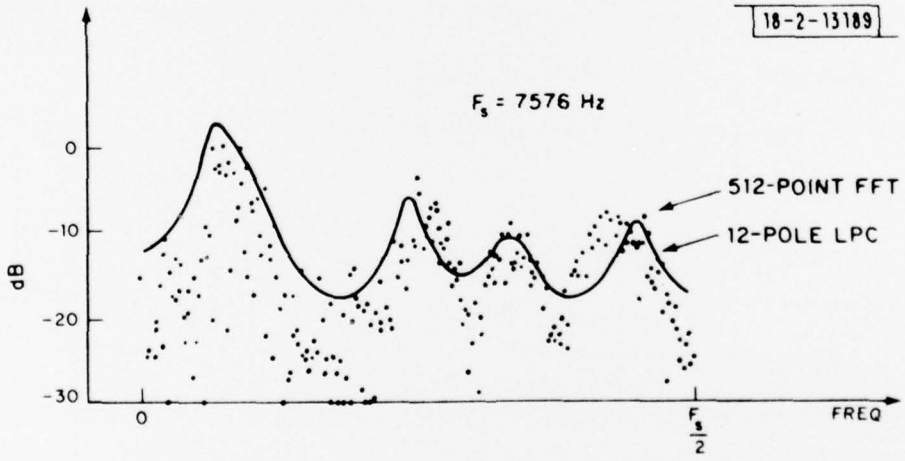


Fig. 10. Voiced speech prefilter output spectra.

a frame represented a mixture of voiced and unvoiced speech the classifier always chose in favour of voiced speech. This event could be reduced significantly by reducing the frame period (10 millisec versus 20 millsec). Although these statistics have been gathered for a relatively small ensemble, the general impression is that the performance is quite good.

Another aspect of the experimental program was the recovery and synthesis of noise-corrupted speech using Linear Prediction techniques. The voiced-unvoiced decisions and the pitch estimates were derived using the methods described in this paper. The LPC filter coefficients were estimated from the prefilter output waveform. For the case of noise-corrupted unvoiced speech, Figure 7a for example, the prefilter output is shown in Figure 7e. Its short term power spectrum is shown in Figure 9 which when compared with that for the input unvoiced speech plus ACP noise, Figure 7f, clearly demonstrates the action of the adaptive prefilter in eliminating the clutter. The LPC power spectrum estimate is also plotted on Figure 9 and shows that the synthetic speech is likely to reproduce the original unvoiced speech. Of course the ACP noise will cause the spectral estimate to be somewhat distorted but the perception of the additive ACP noise will have disappeared. It is for this reason that the synthetic speech is perceived to be "noise-free".

Similar results are obtained for the voiced speech sample function shown in Figure 8a. The short-term power spectrum of the prefilter output, Figure 8e, is plotted in Figure 10 and should be compared with the voiced speech plus noise power spectrum shown in Figure 8f. The corresponding

LPC spectrum shown in Figure 10 shows the distortion in the first format due to the presence of the ACP noise.

LPC synthetic speech was generated for a number of utterances recorded in ACP noise. Compared to LPC speech in which no adaptive prefiltering was employed, an improvement in intelligibility was obtained.

IX. CONCLUSIONS

Using statistical decision theory a new speech classification algorithm has been developed in the form of an estimator-correlator receiver. The structure is robust in the sense that it can adapt to time-varying noise fields in which the signal-to-noise ratio can be quite low (less than 10dB). For noiseless speech the classifier simply involves two fixed filters and requires no pitch estimation or linear prediction analysis parameters. For noisy speech clutter filters must be added to the speech and reference channels. The reference clutter filter is developed on the basis of an initial .3 sec sample of noise data while the other adapts to the speech plus noise statistics calculated for each frame. If a frame is classified as noise, the reference channel filter is up-dated so that time varying noise statistics can be tracked.

The output of the speech channel clutter filter represents an improved estimate of the input speech in the sense that much of the additive noise has been cancelled from the signal. By applying Linear Prediction techniques to this waveform, more intelligible synthetic speech can be obtained.

A relatively thorough (non-real-time) evaluation of the classifier and adaptive prefilter was conducted for Airborne Command Post noise and surprisingly good results were obtained. Based on a limited number of listening tests, the LPC synthetic speech using the prefilter output was found to be more intelligible than the LPC synthesis of the original noisy speech.

No attempt was made to optimize the design of the fixed voiced (low pass) and unvoiced (high pass) filters. In this study 21-tap linear phase filters were used. A better approach would be to obtain long term statistics for voiced and unvoiced speech and pick the filter length and passband edges to more closely represent the average spectral properties. Another useful study would be to investigate the possibility of using recursive filters with phase compensation to further simplify the processing.

Although a first order attempt was made to improve the design of the clutter filters, other methods are undoubtedly possible. Additional insights are also needed in the selection of the clutter filter design parameter; in this note trial and error was used to make the selection.

Of course, the real test of any speech processing algorithm is obtained in a real-time environment. This is the focus of the current effort.

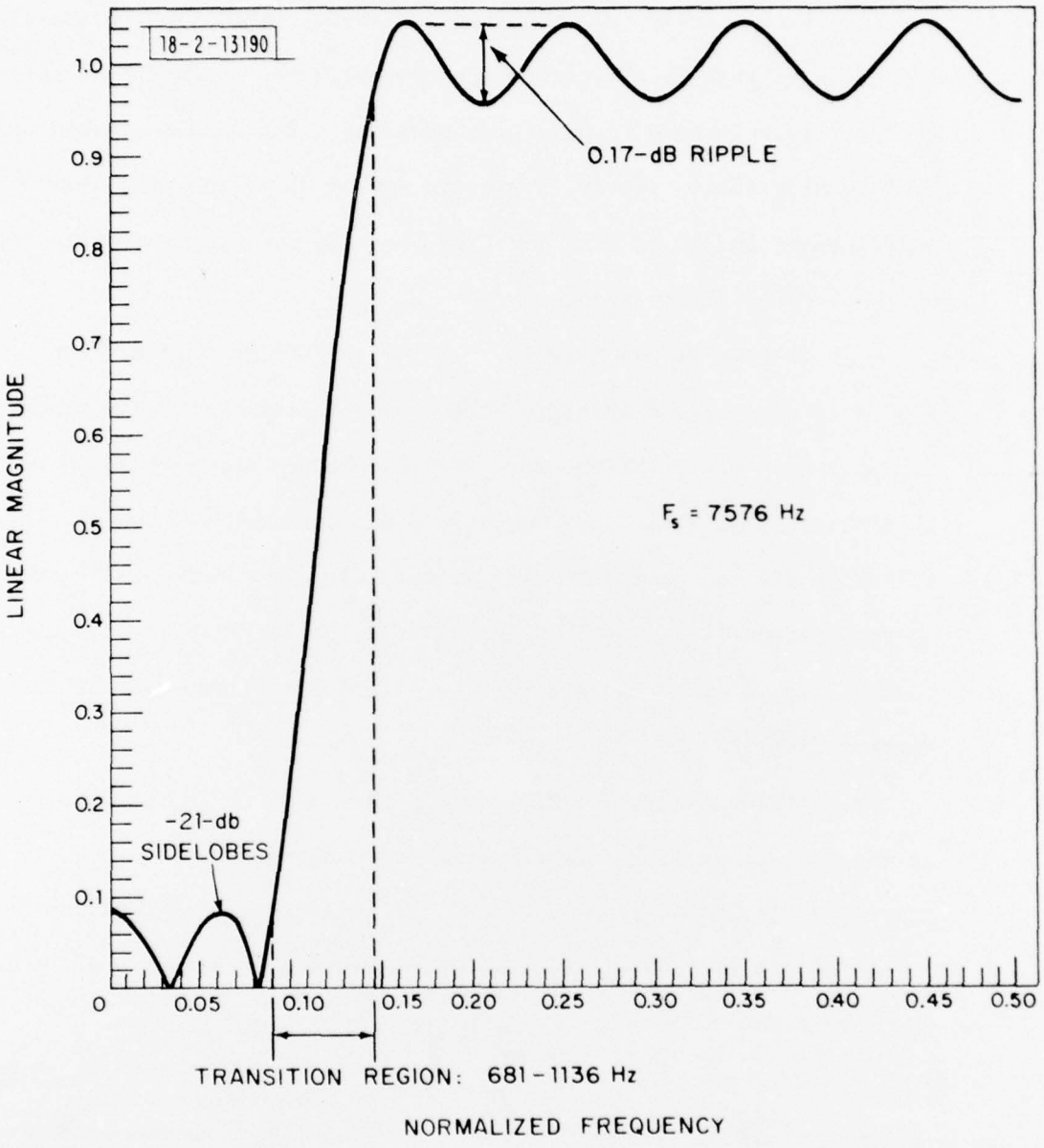


Fig. 11. Unvoiced high-pass filter.

APPENDIX

The unvoiced and voiced speech Wiener filters were approximated by 21-tap linear phase high and low pass filters designed using the Parks-McClellan algorithm. The impulse responses used in the experimental program are given in Table 2 (h(n) = h(-n)). The magnitude of the frequency responses are shown in Figures 11 and 12.

TABLE 2

IMPULSE RESPONSE	UNVOICED FILTER	VOICED FILTER
h(1)	-0.21511067E-01	-0.38655568E-02
h(2)	0.55939741E-02	-0.32053679E-01
h(3)	0.21661893E-01	0.23418449E-01
h(4)	0.39310634E-01	0.13665602E-01
h(5)	0.45899481E-01	-0.42199165E-01
h(6)	0.29383000E-01	0.73566064E-02
h(7)	-0.15331455E-01	0.66053927E-01
h(8)	-0.82191288E-01	-0.65457523E-01
h(9)	-0.15448785E+00	-0.84543467E-01
h(10)	-0.21035391E+00	0.30347985E+00
h(11)	0.76869851E+00	0.59147525E+00

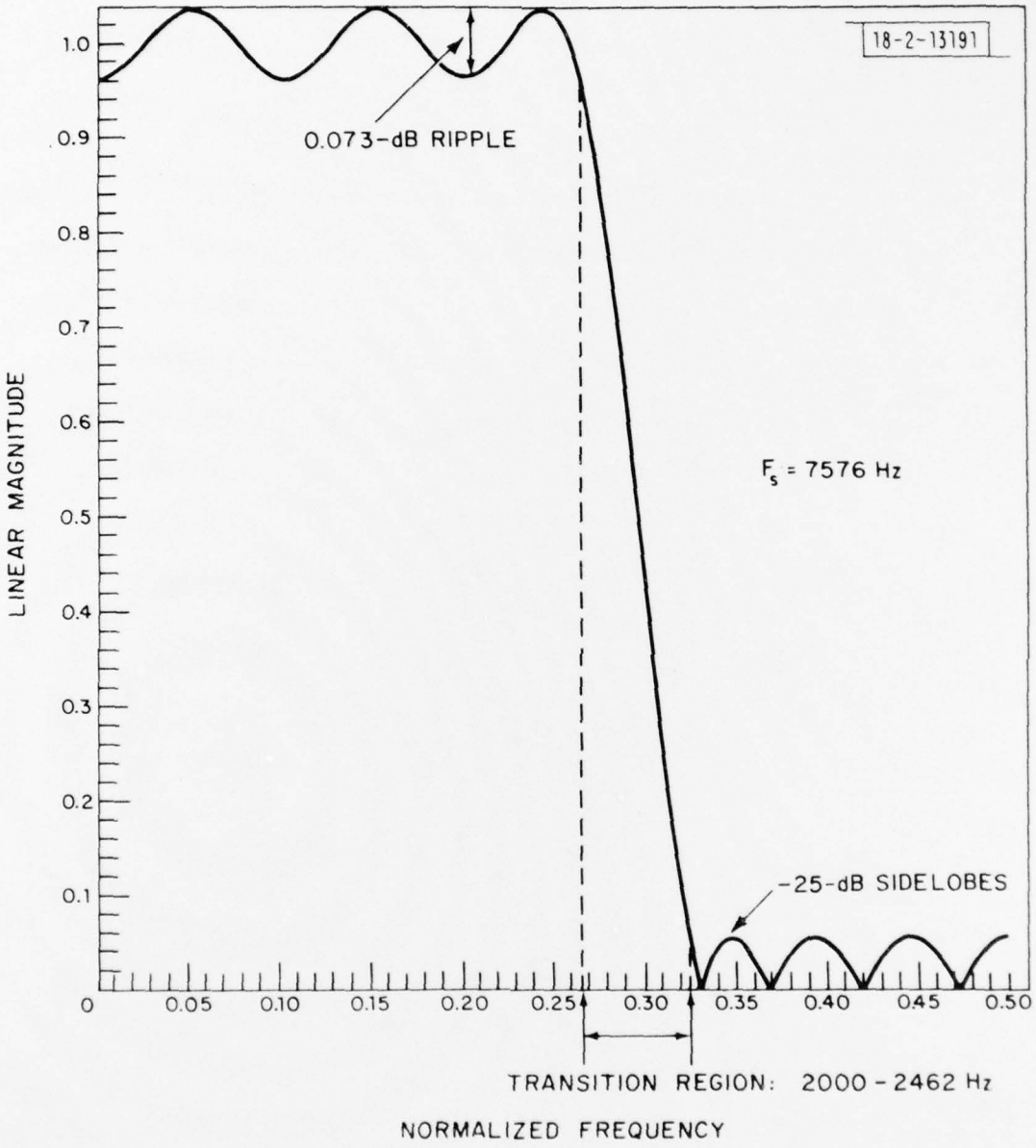


Fig. 12. Voiced low-pass filter.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Ben Gold who highlighted the issues involved in the quest for more robust digital speech terminals. His coaching on the most important properties of speech signals led to the proper formulation of the classification problem. Appreciation is also extended to Dr. Ed Hofstetter, Ms. Stephanie Seneff and Mr. Joe Tierney who listened patiently to premature presentations of the work and who politely guided the author into a more realistic understanding of the speech processing problem. Finally it is noted that the experimental results would not have been possible without the assistance of Dr. Hofstetter and his generous contribution of many I/O software routines.

REFERENCES

1. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. 50, 637-655 (1971).

2. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE 63, 561-580 (1975).

3. J.D. Markel and A.H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-22, 124-134 (1974).

4. B.S. Atal and M.R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell System Tech. J. 49, 1973-1986 (1970).

5. B. Gold, "Robust Speech Processing," Technical Note 1976-6, Lincoln Laboratory, M.I.T. (27 January 1976), DDC AD-A021899/0.

6. B.S. Atal and L.R. Rabiner, "A Pattern-Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition IEEE Trans. Acoust., Speech, and Signal Processing ASSP-24, 201-212 (1976).

7. H.L. Van Trees, Detection, Estimation and Modulation Theory, Part III (Wiley, New York, 1968).

8. H.L. Van Trees, Detection, Estimation and Modulation Theory, Part I (Wiley, New York, 1968).

9. N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction," J. Math. Phys. 25, 261-278 (1947).

10. J.D. Wise, J.R. Caprio, and T.W. Parks, "Maximum Likelihood Pitch Detection," Technical Report No. 7516, Department of Electrical Engineering, William Marsh Rice University (17 October 1975).

11. R.J. McAulay, "Optimum Classification of Voiced Speech, Unvoiced Speech and Silence in the Presence of Noise and Interference," Technical Note 1976-7, Lincoln Laboratory, M.I.T. (3 June 1976), DDC AD-A028518.

12. J.A. Moorer, "The Optimum Comb Method of Pitch Period Analysis of Continuous Digitized Speech," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-22, 330-338 (1974).

13. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust., Speech, and Signal Processing ASSP-22, 353-362 (1974).

14. R.J. McAulay and R.D. Yates, "Realization of the Gauss-in-Gauss Detector Using Minimum-Mean-Squared-Error Filters," IEEE Trans. Inf. Theory, IT-17, 207-209 (March 1971).

15. J.H. McClellan, T.W. Parks, and L.R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," IEEE Trans. Audio Electroacoust. AU-21, No. 6, 506-526 (December 1973).

EXTERNAL DISTRIBUTION

M. Athans, MIT, Room 35-206
Dr. R.P. Wishner
Dr. L.P. Seidman
Systems Control Inc.
260 Sheridan Avenue, Suite 314
Palo Alto, CA 94306

Dr. R. Esposito
Raytheon Company
28 Seyon Street
Waltham, MA

Dr. H. Urkowitz
General Atronics Corp.
1200 East Mermaid Lane
Philadelphia, PA 19118

Prof. D.J. Sakrison
Dept. Electrical Eng.
University of California
Berkeley, CA 94720

Dr. S. Horing
Math'l Analysis and Consulting Group
Bell Telephone Company
Whippany, NJ

Dr. R. Price
Sperry Rand Research Center
Sudbury, MA

Mr. Branko Leskovar
Building 80-024
Univ. of California
Lawrence Berkeley Laboratory
Berkeley, CA 94720

Dr. V.G. Hansen
Raytheon Company, Box A-24
Equipment Development Labs
Boston Post Road
Wayland, MA 01778

Dr. S.B. Weinstein
Bell Laboratories
Holmdel, NJ 07733

Dr. C. Cook
Mitre Corporation
Bedford, MA 01730

Mr. W. Lurie
115 Old Field Road
Newton, MA 02159

Dr. Chi-han Chen
Electrical Engineering Dept.
Southeastern Mass Univ.
North Dartmouth, MA 02747

Mr. G.A. Andrews
Airborne Radar Branch
Naval Research Laboratory
Washington, D.C. 20390

Dr. Julian Bussgang
Signatron
Hartwell Ave.
Lexington, MA 02173

ESD/DCWS Mail Stop 22
Hanscom Field
Bedford, MA 01730
Attn: C.P. Smith, C. Walter

Defense Communications Engineering Center
1860 Wiehle Avenue
Reston, VA 22090
Attn: R. Sonderegger (R540),
Major Leon Lake (R740)

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209
Attn: R. Kahn (R730)

Bell Telephone Laboratories
Murray Hill, New Jersey 07974
Attn: L. Rabiner (Room 2D-533)

Rice University
Houston, TX
Attn: Prof. Tom Parks

Speech Communications Research Laboratory
800 Miramonte Drive
Santa Barbara, CA
Attn: John Markel

Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90291
Attn: D. Cohen

Culler-Harrison, Inc.
150-A Aero Camino
Goleta, CA 93017
Attn: Glenn Culler
Jim McGill

University of Utah
MED 3/60
Salt Lake City, Utah 84112
Attn: Prof. Tom Stockham
Mr. Ali Atashroo
Prof. S. Boll

Stanford Research Institute
333 Ravenswood Avenue
Menlo Park, CA 94025
Attn: Tom McGill

Codex Corporation
15 Riverdale Avenue
Newton, MA 02159
Attn: Dave Forney

Bolt, Beranek, and Newman, Inc.
50 Moulton Road
Cambridge, MA 02140
Attn: J. Makhoul

Network Analysis Corp.
Beechwood, Old Tappin Rd.
Glencove, NY 11542
Attn: De. Ken Schneider

FOREIGN DISTRIBUTION

Dr. D. C. Agarwal
J. K. Institute of Applied Physics
University of Allahabad
Allahabad - 2U-P, INDIA

Prof. M. Noton
Dept. of Electrical Engineering
University of Waterloo
Waterloo, Ontario, CANADA

Prof. M. Blostein
Prof. J. Farnell
Prof. M. Ferguson
Peter Kabal
Dept. of Electrical Engineering
McGill University
Montreal, Quebec, CANADA

Mr. P.J.A. Prinsen
Physisch Laboratorium
Rijksverdedigingsorganisatie TNO
's Gravenhage,
Oude Waalsdorperwet 63
Postbus 2864
GERMANY

Prof. S.S. Haykim
McMaster University
Dept. of Electrical Engineering
Hamilton, Ontario, CANADA

R. P. McLean
Dept. of Electrical Engineering
Queens University
Kingston, Ontario, CANADA

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ESD-TR-76-320
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*: Optimum Speech Classification and Its Application to Adaptive Noise Cancellation

5. TYPE OF REPORT & PERIOD COVERED: Technical Note

6. PERFORMING ORG. REPORT NUMBER: Technical Note 1976-39

7. AUTHOR(s): Robert J. McAulay

8. CONTRACT OR GRANT NUMBER(s): F19628-76-C-0002

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Lincoln Laboratory, M.I.T.
P. O. Box 73
Lexington, MA 02173

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

Program Element No. 33126K

11. CONTROLLING OFFICE NAME AND ADDRESS

Defense Communications Agency
8th Street & So. Courthouse Road
Arlington, VA 22204

12. REPORT DATE: 9 Nov[illegible]1976

13. NUMBER OF PAGES: 62

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*

Electronic Systems Division
Hanscom AFB
Bedford, MA 01731

15. SECURITY CLASS. *(of this report)*: Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

speech classification
noise cancellation
estimator-correlator
clutter filters
estimator filters
Wiener filter
LPC techniques

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Statistical decision theory is used to develop a strategy to determine whether a given interval of noisy speech data should be classified as voiced speech, unvoiced speech or silence. The classifier takes the form of an estimator-correlator receiver. The estimator filters for unvoiced and voiced speech are well approximated by fixed high pass and low pass filters preceded by adaptive clutter filters that cancel the noise. The noise statistics and clutter filters are updated during the classified silent intervals.

In addition to solving the problem of detecting buzz or hiss in noise, the receiver can be used to prefilter the noisy speech signal prior to vocoding. In particular, the noise-stripped speech signal at the output of the clutter filter was used as the input to a linear prediction vocoder. Based on listening tests, a significant reduction in the noise level was achieved with a corresponding improvement in intelligibility.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

207 650